



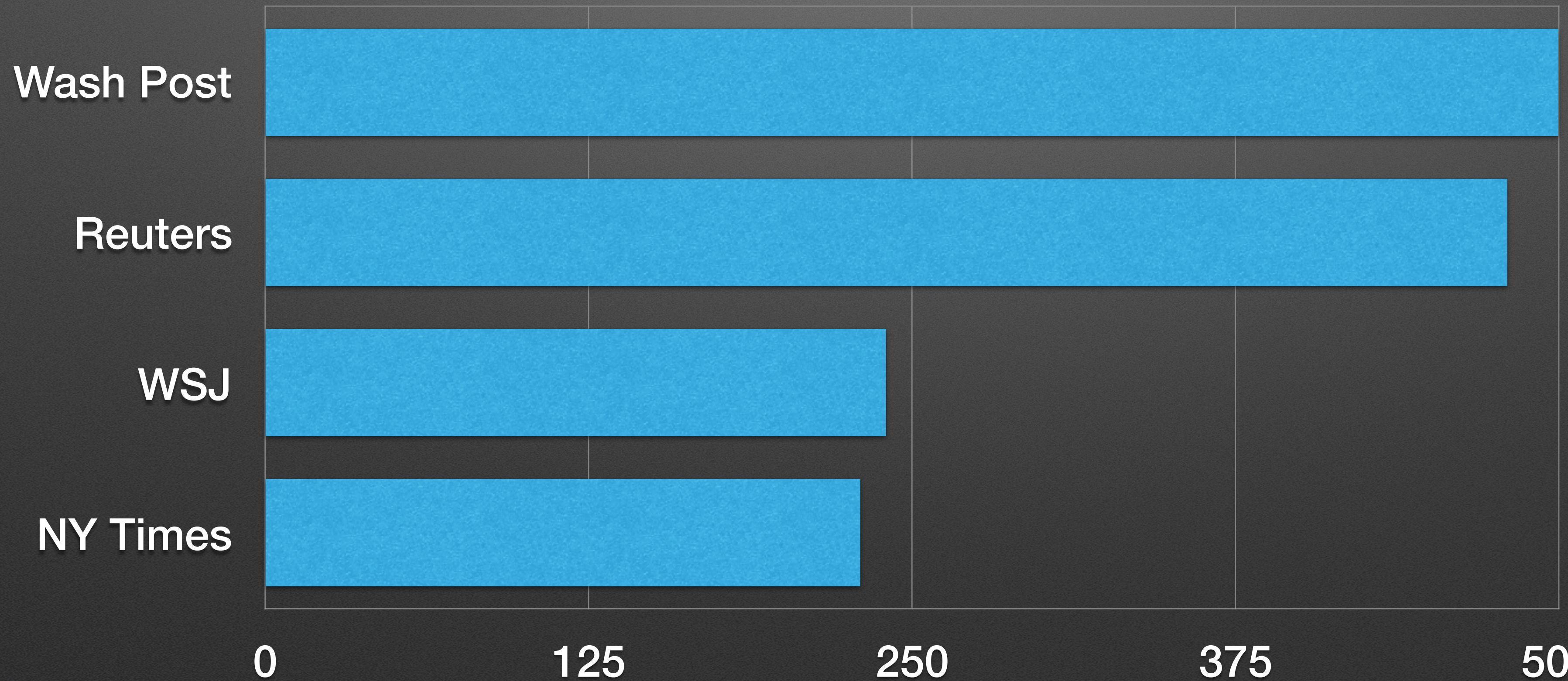
Auto-Gist the News

With Topic Modeling and Extractive Text Summarization Methods

March 2, 2018

Background

Stories Published Per Day (Articles + Video)



The Data

30,000
Reuters News Articles
(January 1, 2018 ~ Present)

- Article URLs pulled using News API (contains links to articles from over 5,000 news sources and blogs)
- Scrapy / BeautifulSoup for scraping content

REUTERS EDITION: U.S. SIGN IN | REGISTER [Twitter](#) [Facebook](#) [RSS](#) [Search News & Quotes](#)

HOME BUSINESS MARKETS WORLD POLITICS TECH OPINION BREAKINGVIEWS MONEY LIFE PICTURES VIDEO

G20 pledges lift Green Climate Fund towards \$10 billion U.N. goal

BY ALISTER DOYLE, ENVIRONMENT AND CORRESPONDENT
OSLO | Sun Nov 16, 2014 1:49pm EST

[Tweet](#) 78 [Share](#) 4 [Share this](#) 8+1 4 [Email](#) [Print](#)



Japan's Prime Minister Shinzo Abe (R) arrives at the G20 Terminal in Brisbane, November 14, 2014.

CREDIT: REUTERS/G20 AUSTRALIA/HANDOUT VIA REUTERS

RELATED NEWS

[U.S., EU override Australia to put climate change on G20 agenda](#)

[G20 statement on Ebola stops short of financial commitments](#)

[Ukraine and Russia take center stage as leaders gather for G20](#)

[Obama, in latest climate move, pledges \\$3 billion for global fund](#)

[Australia's Abbott wants economic reform to be focus at G20](#)

ANALYSIS & OPINION

[Germany just dodges recession](#)

MOST POPULAR

- 1 [REFILE-Saudi oil policy uncertainty unleashes the conspiracy theorists](#)
- 2 [All 50 U.S. states feel freezing temperatures, four dead in New York](#)
- 3 [Palestinians kill five in Jerusalem synagogue attack](#)
- 4 [Nokia revives the brand with launch of iPad lookalike](#)
- 5 [Keystone XL pipeline bill dies in Senate](#)

PICTURES



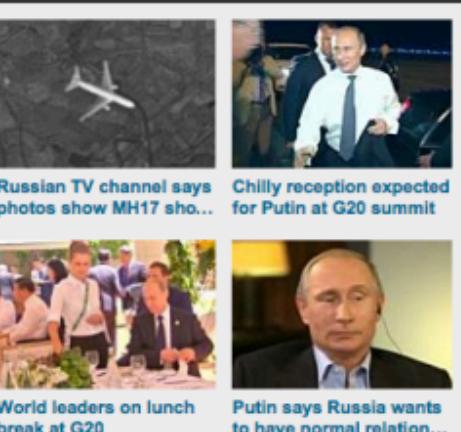
[California's historic drought](#)

With reservoirs at record lows, California is in the midst of the worst drought in decades. [Slideshow](#)

Follow Reuters

[Facebook](#) [Twitter](#) [RSS](#) [YouTube](#)

RECOMMENDED VIDEO



[Russian TV channel says photos show MH17 sh...](#) [Chilly reception expected for Putin at G20 summit](#)

[World leaders on lunch break at G20](#) [Putin says Russia wants to have normal relation...](#)

Topic Modeling

- TF-IDF to reduce weight of terms frequent across documents
- Non-Negative Matrix Factorization (NMF) to extract document topics
- **30 topics total**

Industry-Specific

AIRCRAFT	<i>boeing, airbus, embraer, bombardier, jets</i>
AUTOMOTIVE	<i>gm, vehicles, electric, ford, cars</i>
BUSINESS	<i>percent, billion, quarter, company, revenue</i>
FINANCIAL	<i>bank, banks, billion, financial, funds</i>

Country / Region-Specific (Political)

IRAN	<i>iran, iranian, nuclear, sanctions, tehran</i>
ISRAEL / PALESTINE	<i>israel, israeli, jerusalem, palestinian</i>
NORTH KOREA	<i>north, korea, korean, south, kim, nuclear</i>
SAUDI ARABIA	<i>saudi, arabia, aramco, prince, yemen</i>
TURKEY / SYRIA	<i>turkey, syria, syrian, turkish, ypg</i>

Text Summarization

- 7 Sentence Extraction Algorithms Tested:
 - Luhn
 - Edmundson
 - Lexical Rank
 - Text Rank
 - Sum Basic
 - Latent Semantic Analysis
 - Kulback-Lieber

Luhn Summarizer

- Term frequency determines sentence importance
- TF-IDF for word weighting in document
- Stop word filtering
- Cluster of frequent words indicates good sentence

Reuters) - Chipmaker Microchip Technology (MCHP.O) is in talks to buy Microsemi Corp (MSCC.O), the largest U.S. commercial supplier of military and aerospace semiconductor equipment, a source familiar with the matter said on Tuesday. The company was exploring its options, including a possible sale, after it received a takeover approach, Reuters reported last month. The Wall Street Journal reported earlier Microchip is nearing a deal to buy Microsemi.

Edmundson Summarizer

- Four weighted features for sentence importance:
 - Cue words (e.g. “Significant”, “Greatest”, “Impossible”, “Hardly”)
 - Title & heading words
 - Key word frequency (related to topic)
 - Sentence location

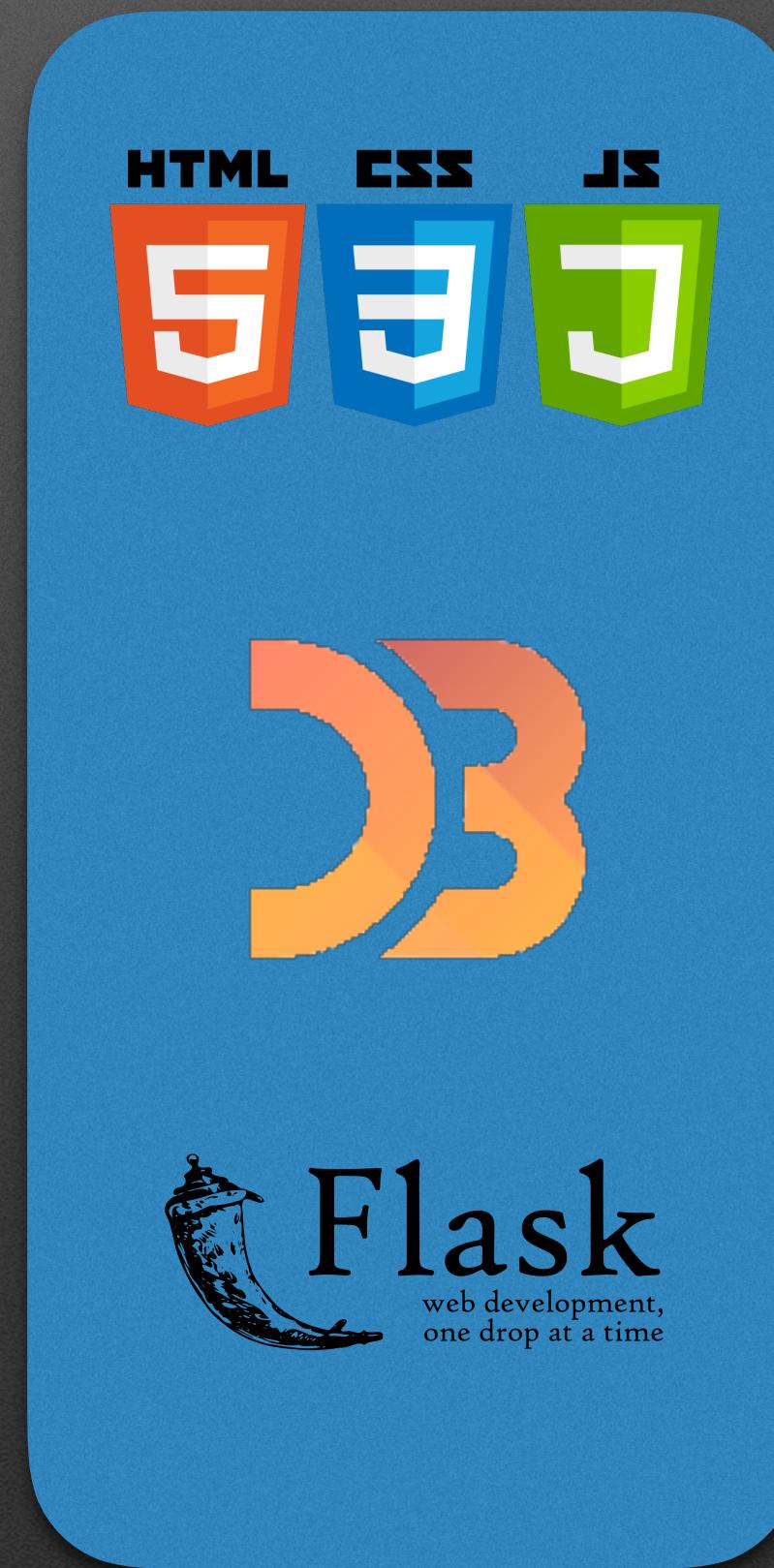
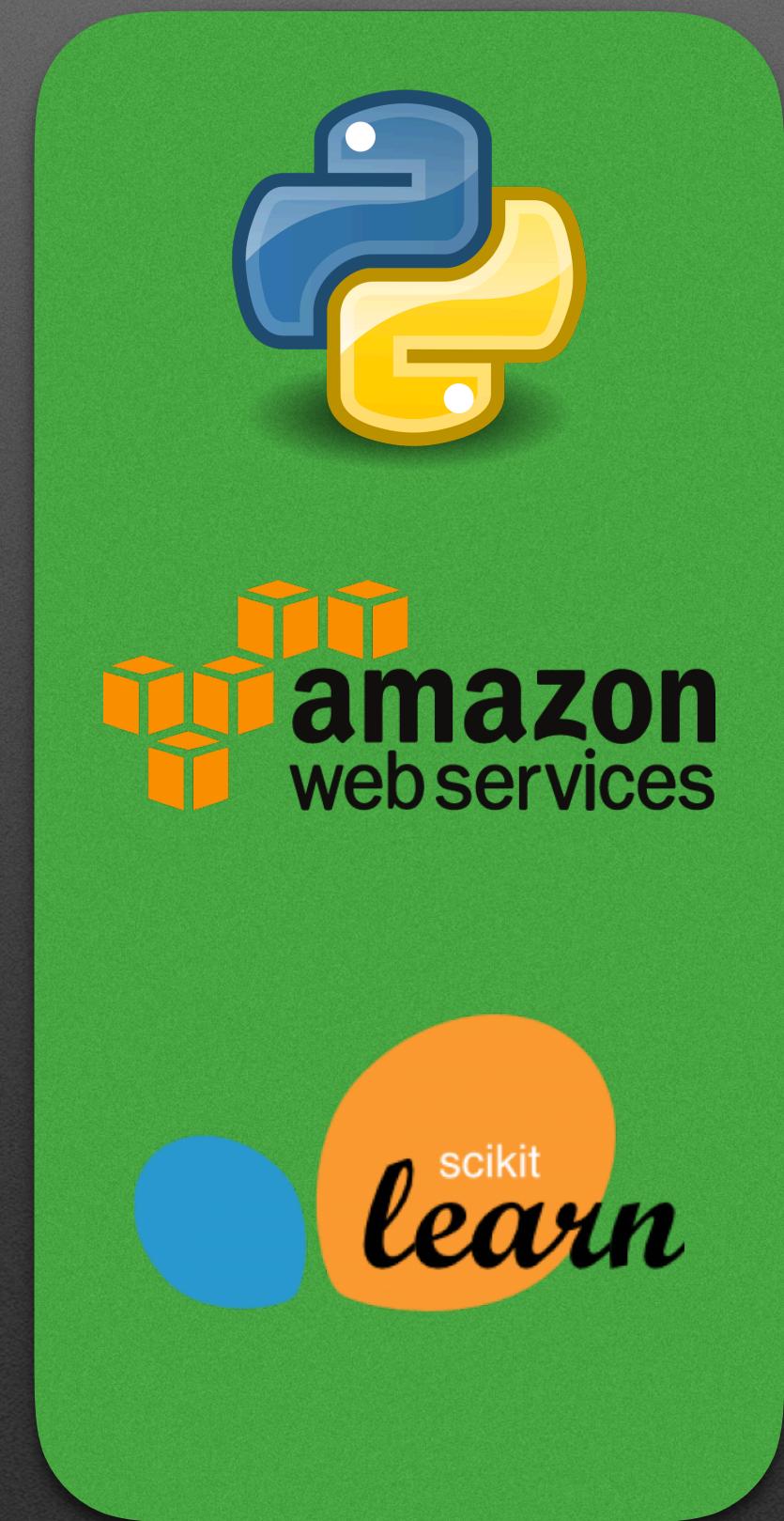
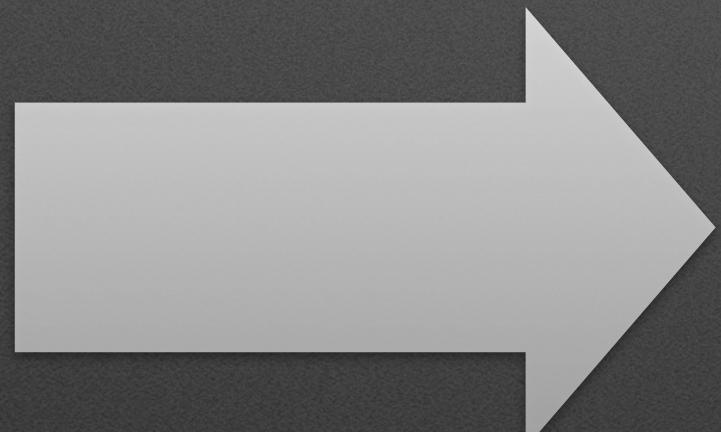
(Reuters) - Chipmaker Microchip Technology (MCHP.O) is in talks to buy Microsemi Corp (MSCC.O), the largest U.S. commercial supplier of military and aerospace semiconductor equipment, a source familiar with the matter said on Tuesday. Broadcom Ltd (AVGO.O) on Monday had called a proposal by U.S. semiconductor peer Qualcomm Inc (QCOM.O) for a new meeting to negotiate an increase to Broadcom's \$117 billion acquisition offer "engagement theater" aimed at dodging a takeover battle. The Wall Street Journal reported earlier Microchip is nearing a deal to buy Microsemi.

Summarization In Practice: YRNWS

Data Collection /
Storage

Back End
Processing

Front End
Interactivity



YRNWS

It's Your News, But Shorter

Date Range

February 24, 2018 - February 24, 2018

Blog / Website

Reuter's (www.reuters.com/articles)



Topics of Interest

Summary Length

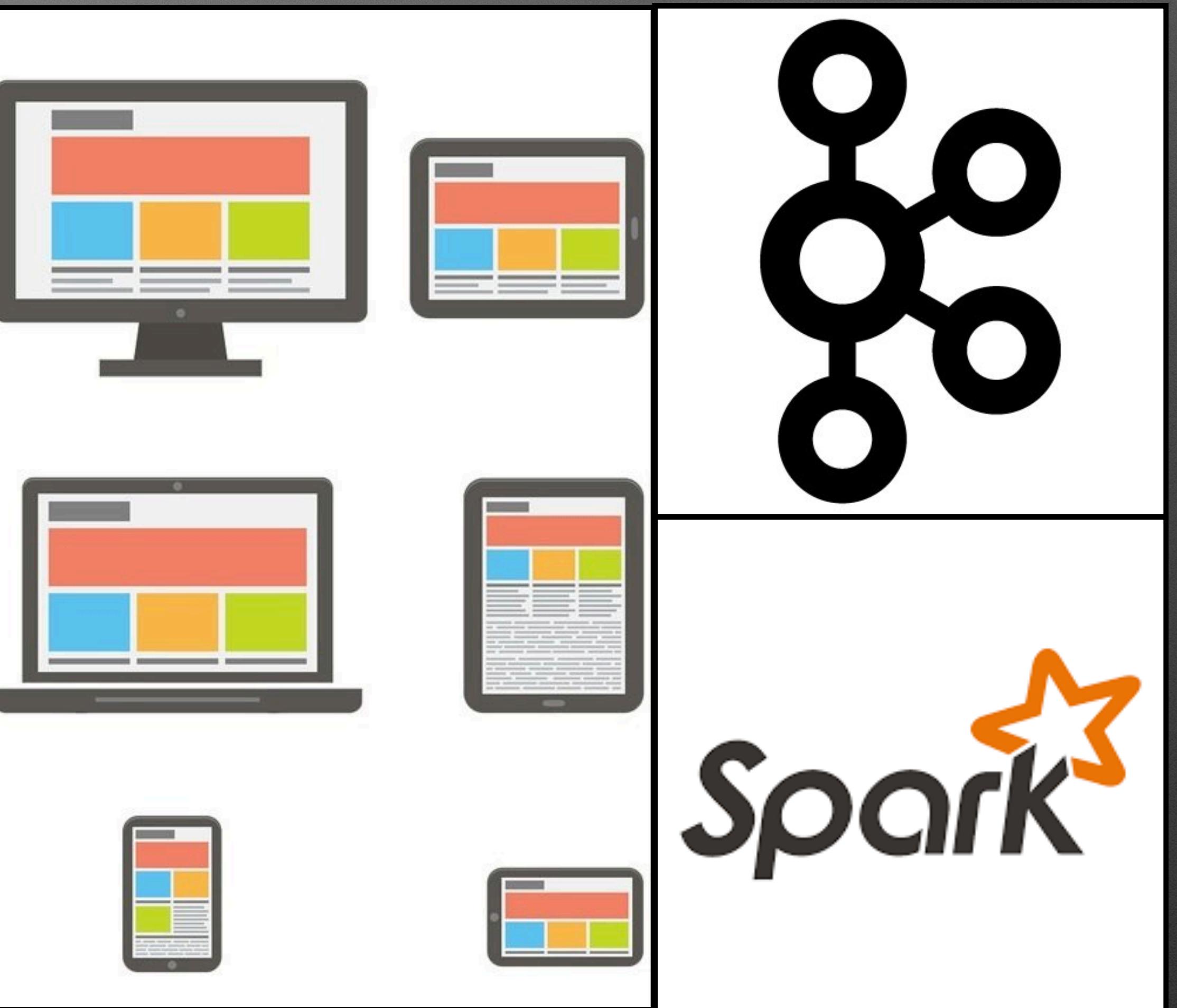
Auto (Currently 3 Sentences)



Generate Report

Next Steps

- Enhanced Data Storage and Streaming
- Web Deployment
- Addition of Other News Sources



Thank You

Topic Modeling

Reuters News

AFGHANISTAN / PAKISTAN	<i>pakistan, afghanistan, taliban, islamabad, afghan</i>	IRAN	<i>iran, iranian, nuclear, sanctions, tehran</i>
AFRICA	<i>zuma, anc, ramaphosa, africa, south</i>	ISRAEL / PALESTINE	<i>israel, israeli, jerusalem, palestinian, palestinians</i>
AIRCRAFT	<i>boeing, airbus, embraer, bombardier, jets</i>	NORTH AMERICA	<i>canada, canadian, nafta, trade, mexico</i>
ASIA	<i>china, chinese, beijing, trade, kong</i>	NORTH KOREA	<i>north, korea, korean, south, kim, nuclear</i>
AUTOMOTIVE	<i>gm, vehicles, electric, ford, cars</i>	OLYMPICS (HIGHLIGHTS)	<i>gold, olympic, medal, team, pyeongchang</i>
BUSINESS	<i>percent, billion, quarter, company, revenue</i>	OLYMPICS (SCANDALS)	<i>doping, athletes, ioc, russian, russia</i>
CRIME / COURT CASES	<i>court, case, supreme, justice, law</i>	SAUDI ARABIA	<i>saudi, arabia, aramco, prince, yemen</i>
ECONOMY	<i>inflation, percent, growth, rate, economy</i>	SECURITY / TERRORISM	<i>police, people, killed, city, attack</i>
ENVIRONMENTAL	<i>oil, crude, bpd, production, opec</i>	SOUTH AMERICA	<i>maduro, venezuela, colombia, opposition</i>
EU (BREXIT)	<i>eu, britain, brexit, european, london</i>	SPORTS	<i>league, game, season, club, team</i>
EUROPE (POLITICS)	<i>party, government, minister, election, parliament</i>	STOCK MARKET	<i>percent, index, stocks, points, dollar</i>
FINANCIAL	<i>bank, banks, billion, financial, funds</i>	TECHNOLOGY	<i>qualcomm, broadcom, apple, nxp, chips</i>
GERMANY (POLITICS)	<i>spd, merkel, coalition, germany, conservatives</i>	TENNIS	<i>match, open, australian, slam, federer</i>
HEALTH / MEDICINE	<i>study, health, patients, women, drug</i>	TURKEY / SYRIA	<i>turkey, syria, syrian, turkish, ypg</i>
IMMIGRATION	<i>myanmar, rohingya, rakhine, bangladesh, refugees</i>	U.S. POLITICS	<i>trump, house, republican, white, democrats</i>